DOCUMENT RESUME

ED 194 583 TM 800 712

AUTHOR Lunneborg, Clifford E.
TITLE Design Coding and Interpretation in Multiple Regression.
INSTITUTION Washington Univ., Seattle. Educational Assessment Center.
PUB DATE Aug 80
NOTE 25p.: A version of this paper was presented at the Annual Meeting of the American Psychological Association (88th, Montreal, Quebec, Canada, September 1-5, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Analysis of Covariance: *Analysis of Variance: Hypothesis Testing: *Mathematical Models: *Matrices: *Multiple Regression Analysis: *Sampling
IDENTIFIERS *General Linear Model

ABSTRACT

The multiple regression or general linear model (GLM) is a parameter estimation and hypothesis testing model which encompasses and approaches the more familiar fixed effects analysis of variance (ANOVA). The transition from ANOVA to GLM is accomplished, roughly, by coding treatment level or group membership to produce a set of predictor or dependent variables. How treatments should be coded is controversial. The present paper redevelops the independence of hypothesis testing and of estimation (point or interval) from the coding of variables. Further, this redevelopment, based on the simple concepts of a design template (a k x k matrix in the k groups design) and of an expectation template (a k element vector), permits mean contrasts or parameter constraints to be directly stated and easily translated into a natural and common computational form. The technique is developed in three parts. First, the essentials of the GLM are reviewed. Then, forms for stating hypothesis or estimating contrasts are developed which will be independent of coding. Finally, the approach is illustrated for the three Serlin and Levin designs, for unbalanced designs, and for designs with missing cells. (Author/CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED194583

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C. Lunneborg

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Educational Assessment Center

University of Washington

August 1980

81-5

Design Coding and Interpretation in Multiple Regression

Clifford E. Lunneborg

Abstract

In recent years progressively more behavioral scientists have adopted the multiple regression or General Linear Model (GLM) as an approach to the design and analysis of studies which would have earlier been interpreted in an analysis of variance (ANOVA) framework. Accompanying the shift has been an appreciable uncertainty about how best to code the design information. It is here demonstrated that the amount of information available for the GLM is independent of the choice of design coding. Making optimal use of the GLM, however, requires an understanding of the underlying population sampling model. Once this model is accepted most ambiguities of analysis and interpretation disappear.

Project 525: A version of this paper was presented at the 88th Annual Convention of The American Psychological Association, Montreal, September 1980.

Design Coding and Interpretation in

Multiple Regression

Behavioral researchers have become progressively more interested in the multiple regression or general linear model (GLM) as a data analytic technique. This very broadly applicable parameter estimation and hypothesis testing model at once encompasses the more familiar fixed effects analysis of variance (ANOVA) approaches while extending them. The transition from ANOVA to GLM is accomplished, roughly, by coding treatment level or group membership to produce a set of predictor or dependent variables. Just how treatments should be coded, however, is for many researchers a puzzling if not controversial topic.

Mutually exclusive categorical variables, dummy variables and effect codes were among the earliest prescribed techniques with researchers choosing from among them depending upon the presumed importance of denoting the group membership of an observation, of including an intercept in the regression solution or of having the design matrix match the classical linear representation of the ANOVA model. These coding systems have remained popular although there is not complete agreement, for example, as to what constitutes dummy variable coding for a particular ANOVA design. Bolstered by the demonstration of Cohen and Cohen (1975) that hypothesis testing, in the usual ANOVA applications, was unaffected by the choice of coding scheme, GLM advocates were free to code on the basis of personal preference or to use schemes that were handiest given a particular computer program.

The very extended arguments, however, about the appropriate analysis of unbalanced or nonorthogonal designs have been accompanied by claims that there are right or wrong ways of coding treatment. Kaufman and Sweet (1974) and Wolf and Cartwright (1974), for example, suggest that the

coding of the design matrix should be a translation of the specific hypotheses to be tested. Similarly, Keren and Lewis (1977) and others would have the coding reflect the relative frequencies of treatments in certain unbalanced designs. The felt need to have the design matrix reflect as closely as possible the ANOVA effects can also lead to design matrices which are rank deficient (Marks, 1974) or are elaborately constructed to compensate for missing cells (Bradley, 1968) or are intended to distinguish between additive and interactive assumptions (O'Brien, 1976.) Even in criticizing the development of Kaufman and Sweet, Blair and Higgins (1978) warn that some codings taken in combination with some regression analyses (sequences of restrictions on the regression coefficients) may lead to changing the hypothesis of interest and conclude that effect coding is less robust than advertised.

Trying to make the GLM design matrix look like an analysis of variance, in general, or compensate for an unbalance in the design complicates the analyst's task, unduly restricts the flexibility of the GLM and leads to confusion as to the interpretation of hypotheses that are under evaluation. Bogartz (1975) made the point, probably a bit too cryptically, in responding to Wolf and Cartwright, while Carlson and Timm (1974) provided a full-rank formulation of the linear model which, we propose to re-emphasize, allows an experimenter to focus on the hypotheses in the analysis without giving undue concern to the choice of code. Oversimplifying only slightly, the coding controversies and confusions result from attempting to test ANOVA effects without translating these effects into contrasts among population mean values.

Linear regression models the relations among population means and to make the fullest and most effective use of the GLM this needs to be taken

into account. GLM is a very flexible analytic tool. Heiberger's survey (Note 1) provides a clear indication that regression programs can specify any design while the typical ANOVA program is limited. A greater variety of hypotheses may be phrased in terms of contrasts among means than are possible in terms of ANOVA effects. And, to repeat Cohen's message (Cohen and Cohen, 1975; Cohen, 1978) not only may categorical and interval independent variables both be included but the scaling of either type is essentially arbitrary. If hypotheses are phrased in terms of the conditional expectations of the dependent variable, choice of coding is irrelevant to tests of those hypotheses. (Choice between categorical or interval measures as Humphreys and Fleishman, 1974, argue is not arbitrary but speaks directly to the question of what populations have means of interest.)

Failure to develop hypotheses in terms of expectations can lead to problems not only in fixed effects ANOVA designs but in the analysis of studies including categorical and measured predictors and these problems may relate not only to hypothesis testing but to the understanding of the size of an effect (Heiser and Hocking, Note 2).

One of the touted advantages to GLM is that it permits the researcher to get beyond tests of significance, to begin to assess the magnitude of effects. In looking at multiple regression from this latter point of view Serlin and Levin (Note 3) have argued anew for a proper choice of coding for the levels of a qualitative variable. This time the choice is to be made (actually computed) on the basis of establishing a correspondence between the regression parameters and estimates (contrasts or comparisons) of interest. They illustrate their technique for randomized groups designs,

the analysis of covariance and the investigation of treatment by covariate interactions.

Because the coding schemes proposed by Serlin and Levin (as with many of the earlier ones derived from ANOVA hypothesis testing concerns) are fairly involved to apply and because there seems no assurance that two or more estimates of interest deriving from the same study might not require separate codings of the data the work involved may be considerable. In response the present paper provides a redevelopment of the independence of hypothesis testing and of estimation (point or interval) from the coding of variables. Further, this redevelopment, based on the simple concepts of a design template (a k x k matrix in the k groups design) and of an expectation template (a k element vector), permits mean contrasts or parameter constraints to be directly stated and easily translated into a natural and common computational form.

The technique is developed in three parts. First, the essentials of the GLM are reviewed. Then, forms for stating hypotheses or estimating contrasts are developed which will be independent of coding. Finally, the approach is illustrated for the three Serlin and Levin designs, for unbalanced designs and for designs with missing cells.

The General Linear Model

The GLM, including the sampling assumptions associated with parameter estimation and hypothesis testing can be summarized in the following description:

(1) $$Y \sim N_n [E(Y) - X\beta, D^2(Y) = \sigma^2 I] ,$$

where

Y is an n element vector of criterion or dependent variable observations,

X is an n x p matrix of specified predictor or independent variable values,

β is a p element vector of unknown expectation parameters,

σ^2 is an unknown variance parameter

and

I is the n x n identity matrix.

Expression (1) states that the vector Y is an observation from a multivariate normal distribution with a certain mean (or expectation) vector, E(Y), and a certain variance-covariance (or dispersion) matrix, $D^2(Y)$. Among the consequences to this particular multivariate normality are the following:

- Each dependent variable observation, each Y_i, is drawn from a normally distributed population. (As the vector Y is a multivariate normal observation each element of the vector is univariate normal.)
- Each of the normally distributed populations from which dependent variable observations are sampled has the same variance. ($D^2(Y_i) = [D^2(Y)]_{ii} = \sigma^2$ for i = 1,2,3,...,n.)

- Each of the normally distributed populations from which dependent variable observations are sampled has a mean which is a linear function of the values of the p predictors.
  ($[E(Y)]_i = E(Y_i) = X_{i.}' \beta = \sum_{j=1}^{p} X_{ij}\beta_j$.)
- The n dependent variable observations are sampled independently.
  ($Cov(Y_i, Y_j) = [D^2(Y)]_{ij} = 0.$)

The matrix X is frequently referred to as the design matrix as, in ANOVA applications, the predictor data reflect how the study was designed, how treatments were assigned to observational units. If X is basic or of full rank (if no predictor or column of X is redundant to other columns of X) then the maximum likelihood estimator of the β vector is usually given as

(2) $$\hat{\beta} = (X'X)^{-1}X'Y \quad .$$

This vector will have a sampling distribution. That is, successive replications of the same "experiment," utilizing the same design matrix but sampling independently from the contributing normal populations of criterion observations, will yield different estimates of β. Under the assumptions given for the GLM this sampling distribution is stated as:

(3) $$\hat{\beta} \sim N_p[\, E(\hat{\beta})=\beta \;,\; D^2(\hat{\beta})= \sigma^2(X'X)^{-1} \,]$$

Over samples the estimator for each element of β is normally distributed about the true value of that parameter.

The maximum likelihood estimator of σ^2 is

(4) $$\hat{\sigma}^2 = (Y - X\hat{\beta})' (Y - X\hat{\beta}) \quad (1/n).$$

This estimator, unlike $\hat{\beta}$, is not unbiased. An unbiased estimator of σ^2, however, is given by

$$\tilde{\sigma}^2 = [1/(n-p)]\ (Y - X\hat{\beta})\ '\ (Y - X\hat{\beta}).$$

Again, because of sample-to-sample uncertainty about the elements of the Y vector the estimators of σ^2 will have a sampling distribution. Owing to the normality of the Y observations these sampling distributions will be related to that of a χ^2 variable:

(5) $(1/\sigma^2)(Y - X\hat{\beta})\ '\ (Y - X\hat{\beta}) = (n\tilde{\sigma}^2/\sigma^2) = [(n-p)\tilde{\sigma}^2/\sigma^2] \sim \chi^2[(n-p)\text{ d.f.}]$.

This sum of squared deviations of the criterion observations about their estimated population means,

(6) $\sum_{i=1}^{n} [Y_i - \hat{E}(Y_i)]^2 = [Y-\hat{E}(Y)]\ '\ [Y-\hat{E}(Y)] = [Y-X\hat{\beta}]\ '\ [Y-X\hat{\beta}] = e'e$,

where $\hat{E}(Y) = X\hat{\beta}$, and $[Y-X\hat{\beta}] = e$, is distributed across samples independently of the distribution of the estimators of β and, thus, we have a way of testing hypotheses about β or of establishing interval estimates of β . Hypothesis testing in the GLM typically takes the form of evaluating the effect of imposing one or more <u>linear constraints</u> on the β vector. If C and Γ are, respectively, a p x k ($< p$) matrix of rank k and a k element vector, both of specified constants, then the relation

(7) $$C'\beta = \Gamma ,$$

imposes k constraints on the p elements of β .

Frequently, Γ is the null vector with elements of C chosen so as to express the hypothesis that a certain element of β is zero or that a certain difference between β elements is zero. A maximum likelihood estimator of β may be obtained subject to the restriction of (7). Denoting this estimator as $_C\hat{\beta}$, a second sum of squared deviations can be defined:

$$f'f = [Y-X\ {}_C\hat{\beta}]\ '\ [Y-X\ {}_C\hat{\beta}] = [Y-\hat{E}(Y|C'\beta=\Gamma)]\ '\ [Y-\hat{E}(Y|C'\beta=\Gamma)] .$$

The likelihood ratio test of the hypothesis of (7) calls for evaluating the ratio

$$(8) \qquad F = \frac{(f'f - e'e)/\ k}{e'e/(n-p)} \sim F_{(k,\ n-p\ d.f.)}$$

which is distributed under the hypothesis of (7) as F with k and (n-p) degrees of freedom. The difference term in the numerator can be obtained from

$$(9) \qquad f'f-e'e = [\Gamma-C'\hat{\beta}]\ '\ [C'(X'X)^{-1}C]^{-1}[\Gamma-C'\hat{\beta}]\ .$$

Commonly C is a vector (only one constraint is being imposed on β) and the inverse called for is simply the reciprocal of a scalar.

Where C is a vector we may also develop a <u>confidence interval</u> for $C'\beta$. Based on (3) $C'\hat{\beta}$ will have a normal sampling distribution:

$$C'\hat{\beta} \sim N_1[E(C'\hat{\beta})=C'\ \beta=\Gamma\ \ ,\ D^2(C'\hat{\beta})=\sigma^2C'(X'X)^{-1}C]$$

or, equivalently,

$$(10) \qquad (C'\hat{\beta}-\Gamma)/[\sigma^2C'(X'X)^{-1}C]^{1/2} \sim N_1[0,1].$$

As this sample statistic is distributed independently of the variance estimators given in (6) we may form a ratio which will be distributed, over samples, as a member of the family of t random variables:

$$(11) \qquad (C'\hat{\beta}-\Gamma)[(n-p)/(e'e[C'(X'X)^{-1}C])]^{1/2} \sim t_{(n-;\ d.f.)}.$$

Limits for $C'\beta$ or Γ can be found by solving this equation for Γ :

$$(12) \quad \Gamma_{lim} = C'\hat{\beta} \pm t_{[\alpha,n-p]}\ [(e'e[C'(X'X)^{-1}C])/(n-p)]^{1/2}.$$

Here $t_{[\alpha,n-p]}$ is a value of the t distribution with n-p degrees of freedom selected so that

$$Pr.[\ t_{[\alpha,n-p)} < t_{(n-p)}] \ = \alpha\ .$$

That is, $(100-2\alpha)\%$ of the intervals constructed in this fashion will contain the true value $C'\beta$.

Design and Expectation Templates

A particular row of the design matrix, $X_{i.}$, carries the predictor or independent variable information for the corresponding, i^{th}, criterion observation. In representing ANOVA, analysis of covariance and related designs in the GLM the X matrix must carry information about group or treatment assignment. How should that information be represented? Although several schemes have been proposed and championed we shall illustrate for the k groups design what freedom and limitations there are. By assumption, the criterion observations have been sampled from k normal populations, each with a potentially different expectation. We represent the set of population means as a vector, the expectation template.

$$(13) \qquad {}_{T}E(Y) = \begin{bmatrix} E(Y|\text{Pop. } 1) \\ E(Y|\text{Pop. } 2) \\ \\ \vdots \\ \\ E(Y|\text{Pop. } k) \end{bmatrix}$$

One reason for calling this an expectation template is that the n element expectation vector, E(Y), may be created from it. Assume that our design calls for sampling (randomly and independently) n_i times from the i^{th} population, i-1,2,3,...,k. We will have included n_1 observations, for example, with a common expectation, E(Y|Pop 1). If we define a replication matrix L with n rows and k columns,

$$(14) \qquad L = \begin{matrix} \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ \end{matrix} \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{matrix} \\ n_1 \\ n_1+n_2 \\ \\ n_1+n_2+\ldots+n_k=n \end{matrix} ,$$

then the product $L\ {}_T E(Y)$ will be an n element vector, the first n_1 elements of which are $E(Y|\text{Pop. }1)$.

We may not choose, however, to collect our data or conduct our experment in such a manner that the n_1 observations from the first population are all at the top of the Y vector. As the E(Y) vector must have elements ordered in the same fashion as Y itself, we can indicate an arbitrary rearrangement or row permutation of the elements of $L\ {}_T E(Y)$, providing an appropriate ordering as

$$E(Y) = \pi\ L\ {}_T E(Y) \quad , \tag{15}$$

where π is an nxn permutation matrix.

The expectation vector for the k groups design consists of n elements but of only k <u>distinct</u> elements. Looking now at the basic GLM expectation equation,

$$E(Y) = X\beta, \tag{1}$$

we can write this anew in template form,

$${}_T E(Y) = {}_T X\ \beta, \tag{16}$$

and refer to ${}_T X$ as the <u>design template</u>. The motivation for the design template is that each observation from Population 1, for example, should be accompanied by the same predictor information. Just as there are only k distinct elements to E(Y) there will be only k distinct <u>rows</u> to the design matrix. The collection of these distinct rows is the design template. The first row of ${}_T X$ contains the predictor (identifying) information associated with an observation from Population 1.

The full design matrix is built up from the template in the same fashion that E(Y) was formed from its template:

$$X = \pi\ L\ {}_T X \tag{17}$$

with π and L as already defined.

What can we say about ${}_{T}X$ beyond noting that it must have k rows? Because we want the k elements in the expectation vector ${}_{T}E(Y)$ on the left hand side of (16) to be mutually linearly independent, because we don't want to build into our design any dependencies among the k population means, it is possible to show that the β vector of (16) must have k elements and that ${}_{T}X$ must be of full rank. That is, ${}_{T}X$ must be a square basic matrix of order k. What we will suggest in the final sections of this paper is that ${}_{T}X$ may be <u>any</u> k^{th} order square basic matrix. It appears to make no difference how we code our groups or treatments so long as we provide k linearly independent codings.

Contrasts and Constraints

The key thing about the general linear model is that it makes statements about the means of populations. Save for those few instances in which we want to draw inferences about the common population variance our hypothesis testing and estimation has to do with those means and with relations among those means. It is a short-cut, and a potentially troubling one when we are interested in group or treatment effects, to think that the βs, the regression coefficients, are the proper focus of our attention. They are not. Their ordering, their orientations, their magnitudes are functions of the choice of ${}_{T}X$. By concentrating instead on the population means we can avoid ambiguities and misunderstandings.

In a preceding section we indicated that defining a set of <u>linear constraints</u>,

$$C'\beta = \Gamma \quad (7)$$

plays a key role in hypothesis testing or estimation in the GLM. We did not say how such a set of constraints arises. At least in the case of designed studies, where the X matrix includes group or treatment information most constraints should follow from a statement of some intended

restriction or restrictions on the means of the populations from which we have sampled. Again it is convenient to have these restrictions in linear form and we can phrase our hypotheses or estimates of interest as contrasts among the population means,

(18) $H'_T E(Y) = \Gamma.$

Each row of H' is a contrast among the k group means and the elements of the row may be chosen, for example, to express the hypothesis that two populations have a common mean or, relatedly, to estimate the magnitude of the difference between two population means. Any non-null k element vector may constitute a row of H'.

Note that by pre-multiplying both sides of (16) by H' we can establish a relation between a set of contrasts and a set of constraints. Corresponding to the contrast

(18) $H'_T E(Y) = \Gamma$

is the constraint

(7) $C'\beta = \Gamma$

if C' is taken to be

(19) $C' = H'_T X.$

Having phrased H' and knowing ${}_T X$ we can calculate, usually very easily, the C' we need for many routine GLM computations.

We can now use our notions of templates and contrasts to look at the three classes of designs discussed by Serlin and Levin (1979), at unbalanced factorial designs and at designs with missing cells. To what extent is the analysis of such designs contingent upon a choice of coding of population sources?

k Randomized Groups

For this design (which includes factorial designs where $k = n_A n_B n_C$, the product of the number of levels of each factor) the expectation vector and design matrix are given completely as

(15) $$E(Y) = \pi L {}_T E(Y)$$

and

(17) $$X = \pi L {}_T X \quad .$$

The only "information" about an observation is the identity of the population, among k, from which it was sampled. Hypotheses can be expressed as

(18) $$H' {}_T E(Y) = \Gamma = C'\beta$$

or a contrast to be estimated given as $H' {}_T E(Y)$.

The test of hypothesis is given in (8) and the confidence interval for the contrast in (12). Doing some calculations based on (15) and (17) we find

$$X'X = {}_T X' L' \pi' \pi L {}_T X$$

or

$$X'X = {}_T X' L' L {}_T X \quad ,$$

as the permutation matrix is an orthonormal matrix, $\pi'\pi = I$,

or

$$X'X = {}_T X' D {}_T X$$

as

(20) $$L'L = \begin{bmatrix} n_1 & 0 & 0 & \cdots & 0 \\ 0 & n_2 & 0 & \cdots & 0 \\ . & . & . & \cdots & . \\ 0 & 0 & 0 & \cdots & n_k \end{bmatrix} = D$$

The inverse of (X'X) can be written

(21) $$(X'X)^{-1} = {}_T X^{-1} D^{-1} ({}_T X')^{-1}$$

as ${}_TX$ is assured of having a regular inverse and, from (2), $\hat{\beta}$ is given as

$$\hat{\beta} = (X'X)^{-1}X'Y = {}_TX^{-1}D^{-1}L'\pi'Y \ .$$

The product $D^{-1}L'\pi'Y$ is simply a vector of the sample means for the k groups,

(22) $${}_T\hat{E}(Y) = D^{-1}L'\pi'Y \ .$$

This estimated expectation template does not depend upon ${}_TX$.

$\hat{\beta}$ can then be expressed as

(23) $$\hat{\beta} = {}_TX^{-1}{}_T\hat{E}(Y) \ .$$

The estimated expectation vector, $\hat{E}(Y)$, based on (17) and (23) is given as

(24) $$\hat{E}(Y) = X\hat{\beta} = \pi L {}_T\hat{E}(Y) \ .$$

Note that $\hat{E}(Y)$ does not depend upon our choice of ${}_TX$.

Since E(Y) is independent of our coding choice so too will be e'e, the error sum of squares for the likelihood ratio test of our hypothesis:

(6) $$e'e = [Y - E(Y)]'[Y - E(Y)]$$

The estimator of the contrast, $C'\hat{\beta}$, can be expressed from (19) and (23) as

(25) $$C'\hat{\beta} = H' {}_T\hat{E}(Y) \ .$$

Again, this estimator is independent of any treatment or group coding scheme. Further, using (19) and (21) yields

(26) $$C'(X'X)^{-1}C = H'D^{-1}H.$$

From (25) and (26) we see as well that the numerator of the hypothesis test,

(9) $$(f'f - e'e) = [\Gamma - C'\hat{\beta}]' \ [C' \ (X'X)^{-1}C] \ [\Gamma - C'\hat{\beta}] \ ,$$

is insensitive to ${}_TX$. In short, testing hypotheses which are in the form of contrasts among the k population means proceeds independently of any specification of ${}_TX$.

Similarly, the confidence interval for the contrast $H'_T E(Y)$, as given in (12), can be shown to be uninfluenced by choice of design codes. Neither hypothesis testing nor contrast estimation depends upon how the design variates in the predictor matrix have been coded.

Analysis of Covariance

In this design, perhaps better known to some readers as the concomitant variable design, the GLM expectation equation may be taken to have the following realization:

$$(27) \qquad E(Y) = X\beta = [Z, L\ {}_TX] \begin{bmatrix} {}_Z\beta \\ D\beta \end{bmatrix} = Z\ {}_Z\beta + L\ {}_TX\ {}_D\beta \ .$$

That is, we have the basic k randomized groups design but associated with each Y observation is not only the group to which it belongs but "scores" on one or more concomitant (predictor) variables or covariables. The design matrix is partitioned between p columns of covariate information and k columns of group or treatment information. The β vector is conformably partitioned. (The matrices L and ${}_TX$ are as described for the k groups design. We omit permutation matrices as would be necessary to mix either observations from the different groups or covariates with design variables. While complicating the algebra, permutations of rows or columns of the design matrix would not affect the outcome of the following developments.)

From this representation we can construct an expectation template, the k element vector, ${}_TE(Y) = D^{-1}\ L'\ E(Y) = D^{-1}\ L'\ Z\ {}_Z\beta + {}_TX\ {}_D\beta$ or

$$(28) \qquad {}_TE(Y) = \bar{Z}\ {}_Z\beta + {}_TX\ {}_D\beta$$

where $\bar{Z}$ is a k x p matrix of mean scores on the p covariates for the k groups. From this we can define an adjusted expectation template,

$$(29) \qquad {}_TE(Y.Z) = {}_TE(Y) = \bar{Z}\ {}_Z\beta = {}_TX\ {}_D\beta \ .$$

Hypotheses of interest in the concomitant variable (covariable) design are almost invariably contrasts among these adjusted expectations:

$$(30) \qquad H'\ {}_T E(Y.Z) = H'\ {}_T X\ {}_D\beta = [0',\ H'\ {}_T X] \begin{bmatrix} Z^\beta \\ D^\beta \end{bmatrix} = C'\beta \ .$$

That is, it is the design betas, rather than the covariate betas that are to be constrained.

Although the algebra is not quite as neat as for the k groups design one can find the inverse of the partitioned matrix

$$X'X = \left[\begin{array}{c|c} Z'Z & Z'\ L\ {}_T X \\ \hline {}_T X'\ L'\ Z & {}_T X'\ D\ {}_T X \end{array}\right]$$

and from that show that $C'(X'X)^{-1}C$, $C'\hat{\beta}$ and $\hat{E}(Y)=X\hat{\beta}$ are each independent of choice of ${}_T X$. One representation of these solutions is

$$(30) \qquad C'(X'X)^{-1}C = H'\ K\ H \ ,$$

$$(31) \qquad C'\hat{\beta} = H'\ K\ L'\ [I - Z(Z'Z)^{-1}Z']\ Y$$

and

$$(32) \qquad X\hat{\beta} = \{ Z(Z'Z)^{-1}Z' + [I-Z(Z'Z)^{-1}Z']\ LKL'\ [I-Z(Z'Z)^{-1}Z'] \}\ Y$$

where

$$(33) \qquad K = \{ L'\ [I-Z(Z'Z)^{-1}Z']D \}^{-1}.$$

(Equations (30) to (33) are certainly not intended as computational forms. They merely illustrate the independence of these matrix terms, as routinely computed in a GLM solution, of the choice of coding.)

These results are sufficient to establish that, as in the k groups design, e'e and (f'f-e'e) are independent of ${}_T X$. Again the test of

hypotheses and the determination of confidence intervals for contrasts among the adjusted means is not contingent upon any particular representation of treatment or group assignment.

Treatment by Covariate Interactions

This design is rather more like the classical analysis of covariance design than the one just discussed in that it permits evaluation of the separate within group regressions of the dependent variable on the covariates. S rlin and Levin (1979) discuss this design in the context of aptitude by treatment interaction studies.

The GLM expectation equation is of the form

$$(34) \qquad E(Y) = X\beta = [\,{}_L Z \,|\, L\ {}_T X] \begin{bmatrix} Z1^{\beta} \\ Z2^{\beta} \\ Zk^{\beta} \\ \hline D^{\beta} \end{bmatrix} = {}_L Z\ Z^{\beta} + L\ {}_T X\ D^{\beta}.$$

The matrix ${}_L Z$ differs from the Z of (27). It will have k x p columns and will be patterned:

$$(35) \qquad {}_L Z = \begin{bmatrix} {}_1 Z & 0 & 0 & \cdots & 0 \\ \hline 0 & {}_2 Z & 0 & \cdots & 0 \\ \hline \vdots & \vdots & \vdots & \cdots & \vdots \\ \hline 0 & 0 & 0 & \cdots & {}_k Z \end{bmatrix} \begin{matrix} \\ - n_1 \\ - n_1+n_2 \\ \\ - n_1+n_2+\ldots+n_k=n \end{matrix}$$

$$\begin{matrix} p & 2p & 3p & & kp \end{matrix}$$

Submatrix ${}_1 Z$ contains scores for the group 1 sample on the p covariates. The

$_{Z}\beta$ vector is partitioned to conform to the columns of $_{L}Z$.

The corresponding expecation template can be written as

$$_{T}E(Y) = D^{-1} L' E(Y) = D^{-1}L'_{L}Z_{Z}\beta + {}_{T}X_{D}\beta$$

or as

(36) $_{T}E(Y) = {}_{L}\bar{Z}' \; Z^{\beta} + {}_{T}X \; D^{\beta}$ where $_{L}\bar{Z}'$ is the k x pk matrix containing vectors of covariate means, one for each of the k groups:

$$(37)\ {}_{L}\bar{Z} = \begin{bmatrix} {}_{1}\bar{Z}' & 0' & 0' & \dots & 0' \\ 0' & {}_{2}\bar{Z}' & 0' & \dots & 0' \\ . & . & . & \dots & 0 \\ 0' & 0' & 0' & \dots & {}_{k}\bar{Z}' \end{bmatrix} .$$

From (36) we can also define, as we did for the concomitant variable analysis, an adjusted expectation template:

$$_{T}E(Y.Z) = {}_{T}E(Y) - {}_{L}\bar{Z}' \; Z^{\beta} = {}_{T}X \; D^{\beta} .$$

There are two classes of hypotheses, contrasts and constraints that are of consequence in the treatment by covariate model. One class consists of contrasts on the adjusted means:

(38) $$H' \; {}_{T}E(Y.Z) = H' \; {}_{T}X \; D^{\beta} = [0' \; , \; H' \; {}_{T}X] \begin{bmatrix} Z^{\beta} \\ D^{\beta} \end{bmatrix} .$$

Are there differences on the dependent variable among the groups (or, how large are those differences likely to be) after we have adjusted for the values assumed by the p covariates? The second class of inferences relates to contrasts among the covariate regressions:

(39) $$H' \; Z^{\beta} = [\, H' \; , \; 0'] \begin{bmatrix} Z^{\beta} \\ D^{\beta} \end{bmatrix} .$$

Is the regression of the dependent variable on one (or more) of the covariates the same within the several groups?

The contrasts among the adjusted means are evaluated as they would be for the concomitant variables design and their independence of the coding of group membership is established in the same fashion. Equations (30) to (33) would be repeated; the Z matrix only would be replaced by ${}_{L}Z$.

By working with the inverse of the partitioned X'X it is possible to show that, in the same fashion, we can obtain for the contrasts among the regression coefficients:

$$(40) \qquad C'(X'X)^{-1} C = H' \{({}_{L}Z'{}_{L}Z)^{-1}[I+{}_{L}Z'LKL\ {}_{L}Z({}_{L}Z'{}_{L}Z)^{-1}]\} H$$

and

$$(41) \qquad C'\hat{\beta} = H'\{({}_{L}Z'\ {}_{L}Z)^{-1}[I-{}_{L}Z'LKL'(I-{}_{L}Z({}_{L}Z'{}_{L}Z)^{-1}{}_{L}Z')]\} Y$$

where, now,

$$(42) \qquad K = \{L'[I-{}_{L}Z({}_{L}Z'{}_{L}Z)^{-1}{}_{L}Z']L\}^{-1} .$$

Similarly $X\hat{\beta}$ is identical to (32), if we simply substitute ${}_{L}Z$ for Z.

The regression contrasts in the treatment by covariate design as well as the contrasts in adjusted expectations are assessed independent of the choice of coding for design. For each of the three standard randomized groups designs, k groups ANOVA, concomitant variable (or covariate) analysis and the analysis of covariance or treatment by covariate analysis the choice of scheme for representing the k groups has no effect on the analysis of the resulting data.

Unbalanced Designs and Missing Cells

The principle which "worked" in the examples above was to concentrate on multiple regression as modelling the sampling of observations from several populations and on formulating hypotheses in terms of contrasts

among the means of those populations rather than in terms of "effects."

This same principle avoids ambiguity in other multiple regression applications. In particular, this is true for the unbalanced factorial design, where different cells are represented in the sample by different numbers of observations. If each cell is taken to be a sample drawn from a different population, contrasts among the population means can be constructed that will correspond to hypotheses of interest.

Where one or more of the cells in the design is represented not just by fewer but by zero observations the analysis of the design, as noted in the introduction, is problematic. If the design is treated as one, however, in which sampling has been from as many different populations as there are populated cells and if one is guided in phrasing hypotheses as contrasts among the means of those populations the resulting analysis becomes clearer.

The GLM provides a powerful approach to design and analysis. To reap the fullest benefits, however, one must keep in mind the sampling model which characterizes multiple regression.

Reference Notes

1. Heiberger, R. M. Conceptualization of experimental designs and their specification and computation with ANOVA programs. Proceedings of the Statistical Computing Section, American Statistical Association, 1976.
2. Heiser, R. F., & Hocking, R. R. On the use of qualitative variables in regression. Paper presented at the 136th annual meeting of the American Statistical Association, Boston, August 1976.
3. Serlin, R. C., & Levin, J. R. Coding for direct interpretation in multiple regression analysis (Occasional paper No. 22). Madison: University of Wisconsin, Laboratory of Experimental Design, 1979.

References

Blair, R. C., & Higgins, J. J. Comments on "Contrast coding in least squares regression analysis." American Educational Research Journal, 1978, 15, 149-151.

Bogartz, W. Coding dummy variables is a waste of time: Reply to Wolf and Cartwright, among others. Psychological Bulletin, 1975, 82, 180.

Bradley, H. E. Multiple classification analysis for arbitrary experimental arrangements. Technometrics, 1968, 10, 13-27.

Carlson, J. E., & Timm, N. H. Analysis of nonorthogonal fixed-effects designs. Psychological Bulletin, 1974, 81, 563-570.

Cohen, J. Partialed products are interactions: partialed powers are curve components. Psychological Bulletin, 1978, 85, 858-866.

Cohen, J., & Cohen, P. Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, N.J.: Erlbaum, 1975.

Humphreys, L. G., & Fleishman, A. Pseudo-orthogonal and other analysis of variance designs involving individual differences variables. Journal of Educational Psychology, 1974, 66, 464-472.

Kaufman, D., & Sweet, R. Contrast coding in least squares regression analysis. American Educational Research Journal, 1974, 11, 359-377.

Keren, G., & Lewis, C. A comment on coding in nonorthogonal designs. Psychological Bulletin, 1977, 84, 346-348.

Marks, E. Analysis of variance techniques for unbalanced data. Review of Educational Research, 1974, 44, 351-364.

O'Brien, R. G. Comment on "Some problems in the nonorthogonal analysis of variance." Psychological Bulletin, 1976, 83, 72-74.

Wolf, G., & Cartwright, B. Rules for coding dummy variables in multiple regression. Psychological Bulletin, 1974, 81, 173-179.